UNIVERSIDADE FEDERAL DO PARANÁ

GUSTAVO ALEXANDRE DUDA MATTANA

ANÁLISE DE CRÉDITO E O *DATA MINING*: UMA PROPOSTA DE APLICAÇÃO NA INSTITUIÇÃO FOMENTO PARANÁ

CURITIBA

2016

UNIVERSIDADE FEDERAL DO PARANÁ

GUSTAVO ALEXANDRE DUDA MATTANA

ANÁLISE DE CRÉDITO E O *DATA MINING*: UMA PROPOSTA DE APLICAÇÃO NA INSTITUIÇÃO FOMENTO PARANÁ

Trabalho de conclusão de curso apresentado como requisito a titulação no programa de Mestrado Profissional em Desenvolvimento Econômico do Departamento de Ciências Econômicas da Universidade Federal do Paraná.

Aluno: Gustavo Alexandre Duda Mattana

Orientador (a): Prof. José Guilherme Silva Vieira

CURITIBA

2016

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS SOCIAIS APLICADAS
Programa de Pós Graduação em DESENVOLVIMENTO ECONÔMICO
Código CAPES: 40001016051P7

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em DESENVOLVIMENTO ECONÔMICO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **GUSTAVO ALEXANDRE DUDA MATTANA**, intitulada: "

"ANÁLISE DE CRÉDITO E O DATA MINING: UMA PROPOSTA DE APLICAÇÃO NA INSTITUIÇÃO FOMENTO PARANÁ"

", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua Aprovação.

Curitiba, 15 de Abril de 2016.

Prof JOSÉ GUILHERME SILVA VIEIRA (UFPR)
(Presidente da Banca Examinadora)

Prof DAYANI AQUINO (UFPR)

Prof JOSÉ WLADIMIR FREITAS DA FONSECA (UFPR)

AGRADECIMENTOS

O conhecimento sobre nossas reais circunstâncias afia nossa visão quanto ao que de fato é relevante na vida, seja de modo individual ou em sociedade. Nosso comportamento se altera quando a visão se amplia, e as circunstâncias, aliadas à um pouco de nossa dedicação, podem então seguir em uma direção mais adequada. Como resultado, nossas vidas melhoram, se aprofundando em alegria e significado. Aos professores, formais e informais que tive até hoje, que são peça chave nesse processo, manifesto minha imensa gratidão. Humildemente, espero um dia poder retribuir ao mundo todo esse auxilio recebido, com juros e correção.

RESUMO

Em um contexto de ampla disponibilidade de informações, a aplicação das metodologias de *Data Mining* provocou, nas últimas décadas, mudanças radicais nos processos de tomada de decisão em diversos campos do conhecimento, dentre eles o das finanças. Uma das beneficiárias desses avanços é a Agência de Fomento do Estado do Paraná - Fomento Paraná, instituição financeira que tem como sua principal atividade a concessão de crédito, orientado a impulsionar o desenvolvimento econômico regional. Tendo como seu acionista majoritário o Governo do Estado do Paraná, a instituição possui amplo potencial de utilização das metodologias de *Data Mining*, uma vez que o mesmo possui em suas bases uma grande quantidade de dados sobre as pessoas físicas e jurídicas do Paraná, sendo que uma considerável parte desses dados contém informações sensíveis ao processo decisório de crédito. Nesse contexto, através da aplicação empírica das metodologias de *Data Mining*, este trabalho teve como objetivo estimar um modelo estatístico básico de análise e suporte a decisão de crédito à instituição financeira de desenvolvimento Fomento Paraná, com foco na utilização de variáveis regressoras cujos valores posteriormente pudessem ser obtidos junto às demais bases de dados administrativas do Governo do Estado do Paraná. Para tanto, foram catalogadas as informações relevantes ao processo decisório de crédito com base na bibliografia acadêmica, e posteriormente foram identificadas bases de dados existentes no Estado que possuem dados e informações dessa natureza. Na sequência, com base no histórico de operações da carteira de Microcrédito da Fomento Paraná, através da aplicação da metodologia de Regressão Logística foi identificado um modelo estatístico básico de analise de crédito, que apresentou graus de Acurácia de até 82%, e que possui um conjunto de variáveis regressoras cujos valores poderão ser acessadas junto as bases de informações administrativas do Governo do Estado do Paraná. Os resultados obtidos permitem que seja estruturado um modelo inicial de análise capaz de agilizar a identificação de empresas com mérito de crédito e dar suporte a tomada de decisão, antes mesmo da instituição ser demandada pelos empreendedores, permitindo que políticas de desenvolvimento regional sejam executadas com maior precisão, agilidade e com a otimização de recursos.

Palavras Chave: Análise de Crédito, Regressão Logística, *Data Mining*, Desenvolvimento Econômico, Bancos Públicos

ABSTRACT

In a context of wide availability of information, where the collection and storage costs are becoming smaller, the process of analysis of credit now has a powerful set of statistical and technological tools - referred as data mining - that radically influenced the agility of decision making process. One of the beneficiaries of this process is the financial institution Agência de Fomento do Paraná – Fomento Paraná, which has as its main activity the concession of credit, aimed to boost regional economic development. However, there is an even greater potential for efficiency gains for the company since it has as its main shareholder the State of Paraná. This occurs because the State Government of Paraná, in the process of providing services to the population, must maintain a large amount of data on individuals and companies of Paraná, and a considerable part of this data contains sensitive information to the credit decision-making process. In this context, by utilizing Data Mining methods, this study aimed to estimate of a statistical model of analysis and credit decision to Fomento Paraná, focusing on the use of variables which the value could later be gathered from the databases of the State Government of Paraná. To this end, sensitive information to the credit decision-making process based on academic literature was cataloged, and existing databases that have information of this nature were later identified in the State. Further, based on the history of the Fomento Paraná Microcredit portfolio, by applying the Logistic Regression method, a basic statistical model analysis of credit that has showed 82% of Accuracy, was identified. The results allow the development of an initial analysis model that permits the identification of companies with credit merit and that can provide support decision-making process, even before the institution is asked by the entrepreneurs, enabling regional development policies to be implemented with greater accuracy, flexibility and resource optimization.

Key words: Credit Analysis, Logistic Regression, Data Mining, Economic Development, Public Banking

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No contexto da história econômica a invenção do crédito representa uma das principais inovações realizadas pela humanidade. Inicialmente permitindo a possibilidade de pequenos ganhos de eficiência e coordenação em economias rudimentares, atualmente consiste em um dos principais pilares das economias contemporâneas, sendo que sua influência no funcionamento das sociedades, mesmo que grande parte despercebida para o grande público demonstra-se cada vez mais relevante.

Porém, desde o início de sua prática a essência do processo de concessão de crédito não se alterou, sendo que ainda consiste fundamentalmente no estabelecimento da confiança de um agente econômico a outro agente econômico, uma tomada de decisão sob um contexto de incerteza. A própria origem da palavra crédito, que tem origem no latim "c*redere",* expressa esse conceito, pois sua tradução significa "acreditar".

Assim, o que era feito com base no relacionamento interpessoal, com a criação dos bancos se sofisticou. Com o surgimento da firma, também a complexidade das interações sociais se intensificou, pois o relacionamento comercial que antes envolvia apenas uma pessoa passou a envolver um conjunto de indivíduos associados, e a responsabilização e a determinação da confiança ficou mais difícil de ser estabelecida. Com o surgimento dessa complexidade, metodologias para a concessão do crédito surgiram, buscando reunir um conjunto de regras práticas para determinar quanto, para quem, e sob quais condições emprestar.

Contudo, ocorridos da historia recente, no advento da Crise de 2008, por exemplo, são uma expressão clara de que o processo de estabelecimento da confiança necessária à atividade do crédito está longe de ser assunto finalizado. Na ocasião, o estopim da crise teve como sua causa principal a concessão de crédito em larga escala a setores da economia americana com dúbia capacidade de pagamento. Este processo contou com amplo respaldo de grandes agências de avaliação de risco, que utilizam metodologias sofisticadas para a avaliação de crédito, e que não impediram grandes prejuízos. A Crise de 2008 também ressaltou

o alto grau de relação ente o crédito e os campos da economia política e do desenvolvimento econômico, e o quanto estes temas ainda carecem de reflexão acadêmica e institucional, uma vez que decisões equivocadas de crédito tomadas em um único país iniciaram um processo que abalou todo o sistema financeiro global, causando uma crise econômica que se demonstra insistentemente presente, mesmo mais de meia década após o ocorrido.

Existe, portanto, amplo potencial de avanço em relação ao tema e justifica-se sua relevância a investigação acadêmica, sendo múltipla sua possibilidade de abordagem. Uma das possíveis abordagens da análise de crédito, foco deste trabalho, guarda relação com um processo histórico que teve seu início em meados do Século XX, e que vêm sendo referenciado como a Terceira Revolução Industrial. Esse período, que se intensificou nas décadas de 1990 e início dos anos 2000, é marcado por um grande conjunto de inovações no campo da ciência da computação e da gestão da informação que estão alterando de maneira radical o processo produtivo de vários setores da economia, e, dentre eles, o setor financeiro.

Nesse contexto, os critérios de confiança subjetiva perderam espaço para metodologias mais técnicas e quantitativas no processo de estabelecimento de critérios para a tomada de decisão. O processo de concessão de crédito passou a ser, em grande parte, parametrizado e automatizado. Além disso, os avanços computacionais no que diz respeito ao armazenamento, a conectividade e o alto poder de processamento habilitaram a criação de quantidades sem precedentes de informações digitais, que passaram a ser utilizadas no processo de concessão. Para a gestão e operacionalização dessa grande quantidade de informações, o setor financeiro tem se aproximado cada vez mais do ainda jovem campo de estudos da ciência da computação chamado de *Data Mining*.

De caráter multidisciplinar, mas com grande ênfase no uso da estatística, o *Data Mining* corresponde à ciência de se extrair conhecimento útil de grandes repositórios de dados, e suas técnicas vêm sendo amplamente utilizadas em setores como a indústria, pesquisa científica, engenharia, e governos e é de consenso geral que este campo de conhecimento guarda um potencial muito grande de impacto para a solução de problemas reais da sociedade (CHAKRABARTI, 2006)

Hoje as atividades bancárias têm maior parte de seu processo decisório amparado por metodologias de tomada de decisão que são formuladas com o

auxílio de ferramental de *Data Mining*, o que proporcionou a indústria consistência nas decisões, agilidade na determinação do crédito, decisões à distância e melhoria na gestão de risco de portfólios. (SICSU, 2009)

Além da indústria bancária tradicional, também se incluem como beneficiárias desse processo as Instituições Financeiras de Desenvolvimento (IFDs), as quais buscam utilizar o Crédito como instrumento facilitador do processo de desenvolvimento das economias regionais. Assim, todo o ferramental do *Data Mining* é passível de ser também utilizado por esta modalidade de instituição, nas quais é estratégica, não só a agilidade na determinação eficiente do crédito, mas também a identificação de quais operações são relevantes a geração de emprego e renda nas regiões.

Em particular no Brasil, existe a figura das Agências de Fomento, que são instituições financeiras com a finalidade de financiamento de empreendimentos com foco no desenvolvimento socioeconômico de suas regiões. Essas instituições possuem atuação regional limitada aos seus Estados, e os Governos Estaduais são seus controladores majoritários. No caso do Estado do Paraná, a Agência de Fomento do Paraná - Fomento Paraná trata-se de uma dessas instituições.

O fato dessas instituições serem controladas pelo poder público regional é uma grande oportunidade pois os Governos Estaduais, em escala e abrangência, estão entre os maiores geradores de dados e informações sobre os indivíduos e empresas em seus territórios, tendo em vista a atuação de suas diversas estruturas institucionais tais como Secretarias de Estado (ex: Saúde, Fazenda, Segurança etc), e Empresas Públicas (ex: fornecimento de água ,energia, transporte, etc.).

À Fomento Paraná é determinante ser eficiente no processo de análise e determinação do Crédito para potencializar o desenvolvimento do Estado. Ao Governo do Estado do Paraná é de interesse o bom funcionamento da economia e, pela natureza de sua atuação junto à sociedade, possui amplas bases de dados que possuem informação sensível aos processos de análise de crédito de indivíduos e de empresas. Nesse contexto, através da aplicação empírica das metodologias de *Data Mining*, este trabalho tem como objetivo estimar um modelo estatístico básico de análise e suporte a decisão de crédito à Fomento Paraná, com foco na utilização de variáveis regressoras cujos valores posteriormente possam ser obtidos junto às demais bases de dados administrativas do Governo do Estado do Paraná.

O trabalho está dividido em quatro capítulos. Este primeiro capítulo consiste na introdução do tema, sua contextualização e apresentação do objetivo geral. O segundo capítulo consiste na apresentação do referencial teórico, onde são apresentados os principais conceitos e metodologias utilizados no processo de análise e concessão de crédito. Também nesse capítulo é discutida a operacionalização destes princípios no contexto do *Knowledge Discovery in Databases* - KDD e do *Data Mining*, com particular ênfase nos métodos de Regressão Estatística e Logística. O terceiro capítulo consiste na etapa de estruturação do modelo de análise de crédito a partir dos referênciais expostos no capítulo anterior. Com base na metodologia KDD, cinco etapas descrevem os passos utilizados na construção do modelo, desde a escolha das variáveis, passando pelo pré-processamento, a transformação, a aplicação do *Data Mining* até a posterior análise dos resultados. O quarto e último capítulo está reservado às conclusões finais do trabalho, onde proposições de futuras ações são apresentadas com base nas evidências levantadas.

2 REFERENCIAL TEÓRICO

Este capítulo contempla a apresentação do referencial teórico utilizado para a realização do trabalho. Primeiramente são expostos os principais referenciais relacionados ao crédito, desde sua conceituação até metodologias de análise e concessão. Essa seção, de caráter mais conceitual, guarda acentuada relevância pois, por explicitar que modalidades de informação são de interesse ao processo de análise de crédito, serviu de amparo para o processo de levantamento de possíveis bancos de dados existentes nas instituições do Governo do Estado do Paraná que contém dados dessa natureza.

Ainda neste capítulo, posteriormente, são expostos os principais conceitos relacionados ao *Data Mining*, bem como seu papel dentro do conjunto de passos do que é referenciado como metodologia KDD, e que serviu como base para estruturação das bases de dados para estimação do modelo exposto nas seções seguintes. Finalizando o capítulo, é apresentado o conjunto de métodos relacionados ao *Data Mining*, com um particular destaque para a Regressão Logística, que foi utilizada para o estabelecimento do modelo estatístico de análise de crédito exposto nos capítulos seguintes.

2.1 O CRÉDITO E O PROCESSO DE ANÁLISE DE CRÉDITO

Sobre a conceituação do Crédito Schrickel (1998) define:

> "Crédito é todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte de seu patrimônio a um terceiro, com a expectativa de que essa parcela volte a sua posse integralmente, após decorrido o tempo estipulado. Essa parte do patrimônio pode estar materializada por dinheiro (empréstimo monetário) ou bens (empréstimo para uso, ou venda com pagamento parcelado, ou a prazo) (SCHRICKEL, 1998, p. 25)"

Crédito é, portanto, a ação de ceder determinado recurso a um terceiro durante um período de tempo determinado, de modo geral mediante a uma remuneração, com base em uma expectativa de retorno. Assim, por sua

conceituação ampla, é natural que o Crédito assuma diferentes formatos, abrangendo atualmente a múltiplas necessidades econômicas tais como: crédito direto ao consumidor, crédito automotivo, cartões de crédito, cheques especiais, crédito estudantil, crédito de capital de giro, crédito para investimentos. crédito para a antecipação de recebíveis etc.

Santos (2003) aponta que dentre as várias possíveis conceituações, uma linha de raciocínio tem predominado entre os autores, na qual o Crédito se refere à troca de um valor presente por uma promessa de reembolso futuro, não necessariamente certa, em virtude do fator de risco, e que a esse processo estão relacionadas duas noções fundamentais: confiança, expressa na promessa de pagamento, e tempo, que se refere ao período fixado entre a aquisição e a liquidação da dívida.

Uma vez que se trata de um processo intertemporal, o crédito figura como um fato econômico de tomada de decisão sobre incerteza. Incapaz de saber sobre todos os possíveis cenários do futuro, ao agente econômico que irá decidir sobre a concessão resta utilizar um modelo mental que lhe sirva como suporte para essa decisão. Essa decisão poderá ser amparada em fatores pessoais do agente ou então poderá ser baseada em dados e informações técnicas, sendo que não há necessariamente um consenso sobre uma metodologia definitiva e que esteja livre de falhas.

Em torno dessa característica do processo de Crédito surgiu todo um campo de estudos referenciado como Análise de Crédito. Sobre a Análise de Crédito Schrickel (1998) define:

> "O principal objetivo da análise de crédito numa instituição financeira (como para qualquer emprestador) é o de identificar os riscos nas situações de empréstimo, evidenciar conclusões quanto a capacidade de reembolso do tomador, e fazer recomendações relativas à melhor estruturação e tipo de empréstimos a conceder, a luz das necessidades financeiras do solicitante, dos riscos identificados e mantendo, adicionalmente, sob perspectiva, a maximização dos resultados da instituição (SCHIRICKEL, 1998, p. 26)"

Nestes termos, a Análise de Crédito consiste em se aplicar alguma determinada metodologia com foco na maximização dos resultados oriundos da atividade de concessão de Crédito, dadas as evidências e circunstâncias envolvidas no processo. Tal tarefa compreende a realização de uma decisão que detenha uma

consistência minimamente lógica em um contexto onde o comportamento de uma série de variáveis determinantes do retorno do crédito são incertas, e as informações disponíveis são assimétricas ou inexistentes.

Sobre os critérios essenciais a serem contemplados na análise de crédito, é consenso entre os autores a necessidade de se averiguar a idoneidade dos clientes (sejam pessoas físicas ou jurídicas) e sua capacidade financeira para amortizar a dívida, e que o conjunto mínimo de informações para que a formação dessa compreensão seja possível são tradicionalmente referenciadas como os 5Cs do crédito. (SCHRICKEL, 1998; SANTOS 2003, GUIMARÃES, 2002, FONSECA et al, 2008)

2.1.1 Os 5Cs do crédito

É comum que cada instituição financeira possua sua própria metodologia de Análise de Crédito que corresponda às suas inclinações ao risco, convicções sobre o negócio, e de acordo com o tipo de público e região onde atuam. Como apontado não há necessariamente um consenso sobre uma metodologia definitiva e que esteja livre de falhas, porém existe um consenso literário em torno de um conjunto de princípios que se utilizados, condicionam as decisão de crédito a uma análise que seja minimante lógica e consistente que são referenciados como os “5 Cs do Crédito”. Os “5 Cs do Crédito” são uma prática tácita amplamente adotada como base para a elaboração de metodologias de análise e das políticas de crédito das instituições financeiras. Os “5 Cs” dizem respeito a 5 dimensões que minimamente devem ser alvo de análise no processo de determinação do crédito. São elas: Capacidade, Capital, Colateral, Condições e Caráter.

2.1.1.1 O conceito de Caráter

O conceito de Caráter do tomador indica uma característica essencialmente subjetiva que trata da vontade do tomador do empréstimo de pagar suas contas,

referindo-se à sua índole, ética e senso moral. O credor irá formar uma opinião quanto a confiabilidade do devedor e sua disposição para pagar o empréstimo ou gerar um retorno sobre fundos investidos em sua empresa (FONSECA et al, 2008).

Mesmo contendo uma característica de subjetividade, é possível identificar de modo razoavelmente claro esse grau de confiabilidade expresso no conceito de Caráter. O credor no exame do caráter de seu cliente, é capaz de realizar uma investigação sobre seus antecedentes, o que comumente é feito mediante a elaboração da ficha cadastral. Através da realização da ficha cadastral, o credor tem acesso às informações básicas da empresa (tais como CNPJ, CNAE, endereço, telefone etc) e sobre a identificação e qualificação dos sócios e administradores (nomes, RG, CPF, profissão etc.).

Em posse das informações cadastrais básicas do empreendimento e de seus sócios é possível traçar um perfil da confiabilidade do mutuário uma vez que através da consulta a base de dados de empresas especializadas em coleta, armazenamento e comercialização de informações relacionadas a idoneidade do cliente no mercado de crédito. Os credores verificam nos arquivos de empresas de gerenciamento de risco de crédito (exemplos Equifax, Serasa) se existem informações desabonadoras dos clientes, tais como decorrentes da existência de ações executivas, cheques devolvidos, protestos, falências, atrasos etc. (SANTOS, 2003).

Dentre exemplos de variáveis relacionadas à dimensão de análise Caráter podem ser citados: elaboração de ficha cadastral do empreendimento contendo nome, CNPJ, CNAE, telefone, email, endereços etc; elaboração de ficha cadastral dos sócios e administradores do empreendimento contendo nome, RG, CPF, profissão, idade, estado civil, escolaridade, endereços, telefone, email etc; consultas a base de dados especializadas para averiguação de aspectos jurídicos tais como protestos, falências etc; consulta as base públicas para emissão de certidões negativas junto a Secretaria da Fazenda Estadual, Receita Federal, etc. consultas as bases de dados do BACEN para identificação do perfil de dívidas do mutuário junto a demais instituições financeiras etc. (SCHRICKEL, 1998; SANTOS 2003, GUIMARÃES, 2002, FONSECA et al, 2008)

2.1.1.2 O conceito de Capacidade

O conceito de Capacidade refere-se ao julgamento do analista quanto à habilidade dos clientes no gerenciamento e conversão de seus negócios em receita. Usualmente os credores atribuem à receita das empresas a denominação de fonte primária de pagamento, sendo este um dos principais referencias para se verificar a condição de honrar a dívida por parte do cliente (SANTOS, 2003).

O conceito busca identificar o nível de maturidade do empreendimento e a possibilidade real de reembolso dos recursos ao credor, através da análise de variáveis que reflitam a eficácia e a eficiência de gestão dos administradores e o próprio grau de especialização da empresa. Dentre exemplos de variáveis que expressam o conceito de capacidade podem ser citados: o tempo de atividade da empresa, nível de experiência dos sócios; nível médio de escolaridade dos empresários e dos demais empregados; existência de sistema para controle de informações gerenciais; controles contábeis e elaboração de Balanços, DREs, Fluxo de Caixa, etc; existência de planos de marketing de produtos e serviços; prática de planejamento estratégico; aspectos relacionados à estrutura organizacional e de pessoas (tais como planos de carreira, estratégias de sucessão, nível de formalização de empregados etc); grau de investimento em pesquisa para o desenvolvimento de novos produtos e processos etc. (FONSECA et al, 2008)

2.1.1.3 O conceito de Capital

O conceito de Capital refere-se ao estudo do patrimônio, da solidez da empresa, ou à estrutura de composição da mesma, no sentido de ter recursos próprios que aplicados na atividade produtiva, geram resultados que permitem arcar com o ônus dos créditos conseguidos junto a terceiros. Assim, para uma análise de crédito, é importante verificar o montante de capital próprio que é empregado em uma empresa e também sua estrutura, comparativamente, ao capital de terceiros,

que deve ser capaz de gerar receita que permita saldar os empréstimos realizados, (GUIMARÃES, 2002)

O conceito de Capital corresponde portanto análise da situação financeira do cliente, levando-se em consideração a composição qualitativa e quantitativa dos recursos, onde são aplicados. Busca identificar se o tomador reúne condições financeiras para honrar seu crédito dentro do prazo estipulado. Assim, de modo geral o conceito de Capital corresponde a uma análise de natureza mais contábil, com foco nos números financeiros do empreendimento, geralmente incluindo uma análise dos seus Demonstrativos Contábeis.

Dentre exemplos de variáveis que expressam o conceito de Capital podem ser citados: o Balanço Patrimonial, composto pelo ativo, passivo e patrimônio líquido da instituição; a Demonstração do Resultado do Exercício que tem como objetivo principal apresentar de forma vertical resumida o resultado apurado em relação ao conjunto de operações realizadas num determinado período pelo empreendimento; Demonstração dos Lucros ou Prejuízos Acumulados, podendo ser substituído pela Demonstração das Mutações do Patrimônio Líquido que demonstram o perfil da evolução patrimonial do empreendimento e sua lucratividade; Demonstração dos Fluxos de Caixa que apontam a saúde financeira da instituição e seu grau de liquidez; Declaração de imposto de Renda Pessoa Jurídica etc. (SCHRICKEL, 1998; SICSU, 2010)

.

2.1.1.4 O conceito de Colateral

O conceito de Colateral é uma tradução do termo em inglês, de idêntica escrita, e significa garantia. Em uma operação de Crédito é necessário que exista algum item tangível que tenha qualidade de reserva de valor que venha a servir de salvaguardas, caso a expectativa de reembolso, que é baseada na análise das demais dimensões, acabe por não se concretizar. (SCHRICKEL, 1998)

Portanto, o conceito de Colateral representa as garantias que estão sendo apresentadas pelo mutuário ao credor como forma de minimizar o risco da operação. As garantias são também item passível de análise, pois existem virtudes e

dificuldades atreladas aos seus diferentes tipos. Uma garantia na modalidade aval é de fácil operacionalização, porém a execução é custosa e tem relativo grau de sucesso. Já garantias de hipotecas de imóveis são excelentes para o credor, porém não são todos os mutuários que possuem essa modalidade de garantia disponível para oferta o que pode impactar no volume de crédito, represando os empréstimos. Há também de se analisar o grau de cobertura da garantia, que pode ser maior ou menor que o valor a ser contratado.

Mesmo que seja ofertada uma excelente garantia por parte do mutuário, nunca essa deve ser o único motivo para se efetuar o crédito, pois somente ela não faz com que o crédito retorne no prazo combinado. Ações judiciais para cobrança do crédito inadimplente podem ser muito demoradas e na maioria das vezes serão questionadas muitas cláusulas do acordo previamente firmado entre as partes. A intenção de qualquer instituição de crédito é o retorno dos empréstimos concedidos e que seja no prazo combinado para que possam ser realizados novos negócios com rendimento melhor que pendências nos tribunais de cobrança de devedores inadimplentes, que podem se arrastar por anos sem solução (FONSECA et al, 2008).

Dentre exemplos de variáveis que expressam o conceito de Colateral podem ser citados: Hipoteca de bens móveis ou imóveis com valor de mercado e adequada liquidez; Alienação Fiduciária de bens e direitos econômicos; Penhor Mercantil; Caução; Ações; Cédula Hipotecária; Certificado de Depósito; Debêntures; Duplicatas; Letras de Câmbio; Letras Hipotecárias; Notas Promissórias; Título de Dívidas; Avalistas ou fiadores; Carta de Fiança; Carta de Crédito; Fundos Garantidores de Crédito etc. (SCHRICKEL, 1998; SANTOS 2003, GUIMARÃES, 2002, FONSECA et al, 2008)

2.1.1.5 O conceito de Condições

O conceito de Condições esta relacionado a sensibilidade da capacidade de pagamento dos clientes à ocorrência de fatores externos adversos ou sistemáticos, tais como os decorrentes de aumento de taxas de inflação, taxas de juros, taxas de

câmbio, sazonalidade, crises econômicas. A atenção a esses fatores é determinante para identificação do grau de risco de Crédito, sendo que ao conceito de Condições também corresponde a análise dos parâmetros da operação, delineando a finalidade de utilização dos recursos, os prazos de carência e amortização, as tarifas adicionais e taxas de juros a serem praticadas etc. (SANTOS, 2003)

Esse conceito engloba portanto uma visão mais ampla do processo de crédito, buscando identificar as principais características do solicitante aos níveis de risco aos quais a instituição esta aberta a negócios, bem como condição de uso a qual disponibiliza seus recursos. Essa adequação das características dos clientes às condições de crédito é referenciada no mercado como “enquadramento”, na qual é realizada uma checagem do perfil dos clientes e sua adequação ou não ao apetite pelo risco da instituição. De modo geral, dada a escala, as condições gerais do perfil de risco da instituição financeira estão refletida no conjunto de produtos de crédito que possui, onde as taxas, prazo, tarifas estão determinadas previamente para determinados tipos de públicos específicos.

Dentre exemplos de variáveis que expressam o conceito de Colateral podem ser citados: Setor de Atuação do mutuário para identificação dos riscos sistêmicos a ele atrelados; Identificação de tendências macroeconômicas e setoriais; grau de dependência das atividades da empresa ligadas ao Setor Publico e Mercados externos; Informações sobre concorrência do setor; Políticas econômicas que possam vir a alterar as condições de comercialização de produtos relacionados com sua operacionalização etc. (GUIMARÃES, 2002, FONSECA et al, 2008)

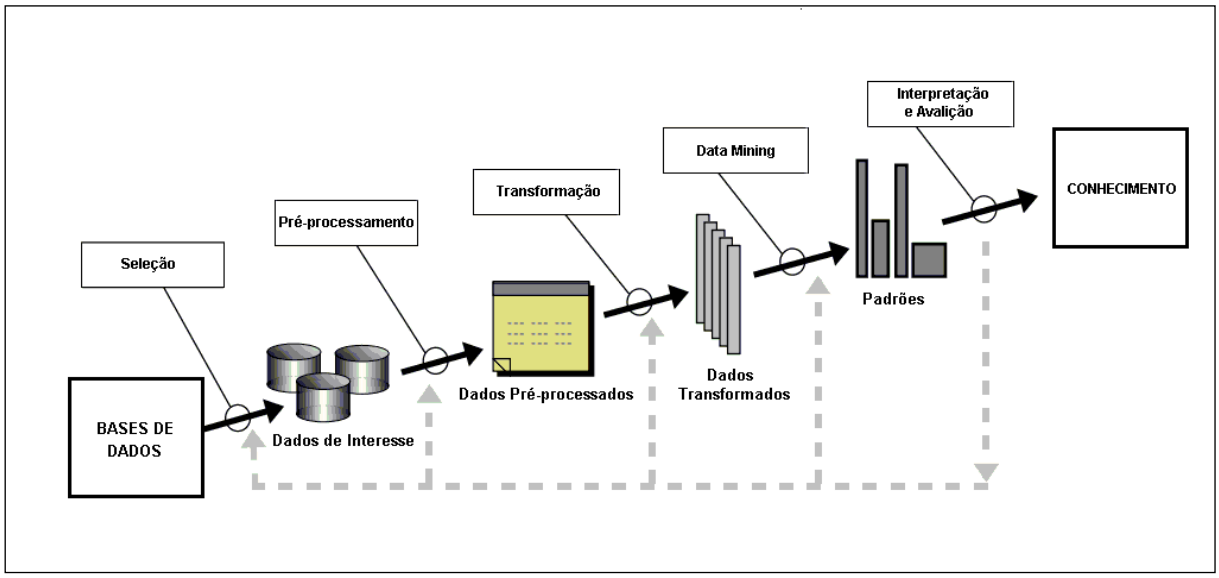
2.2 A ANÁLISE DE CRÉDITO NO CONTEXTO DO *DATA MINING*

Os pontos abordados demonstram a essencial importância das informações ao processo de Análise de Crédito, sendo que o nível correspondente do sucesso das operações de crédito está diretamente correlacionado com o grau de sofisticação da apuração dessas informações. É neste contexto que a Análise de Crédito e o ferramental tecnológico das ciências da Computação se encontram.

De acordo com Lemos (2004) o conjunto de técnicas e ferramentas que buscam transformar os dados armazenados nas empresas em conhecimento é denominando “Descoberta de Conhecimento em Bases de Dados” ou, mais usualmente utilizado, o termo *Knowledge Discovery in Databases* ou KDD.

Segundo Fayyad et al. (1996), o termo KDD foi criado em 1989 como referência ao processo amplo de encontrar conhecimento em dados. KDD refere-se a todo processo de descoberta de conhecimento útil de dados. O autor aponta que até 1995, muitos autores consideravam os termos KDD e *Data Mining* (Mineração de Dados) como sinônimos. Porém atualmente há um consenso quanto à separação entre os termos, porém sendo a etapa de *Data Mining* a etapa mais importante da abordagem KDD, a ser detalhada na sequência. O mesmo autor expõe que a abordagem KDD é composta por 5 etapas: 1 - seleção dos dados; 2 - pré-processamento e limpeza dos dados; 3 - transformação dos dados; 4 - *Data Mining*; e 5 - interpretação e avaliação dos resultados. Essas etapas podem ser visualizadas na figura 1

FIGURA 1 - ETAPAS DO KDD



FONTE: (FAYYAD et al, 2006)

Inicialmente deve-se definir o escopo da informação que se deseja obter para que se inicie o processo de levantamento de bases de dados que contenham dados sensíveis aos objetivos da análise, correspondendo essas ações a etapa de seleção de dados. O segundo passo, correspondente a etapa de pré-processamento

e limpeza dos dados, inicia-se o tratamento dos dados obtidos, buscando adequá-los em termos de formato digital, bem como tratando inconsistências tais como campos vazios, preenchidos equivocadamente, outliers etc. Hand et al (2001) aponta que essa etapa do processo do KDD pode tomar até 80% do tempo necessário de todo o processo, devido às dificuldades de integração de bases de dados heterogêneas. Os dados devem ainda passar por uma etapa de transformação (terceira etapa) visando o armazenamento adequado para aplicação das técnicas da quarta etapa: o *Data Mining.*

De acordo com Fayyad *et al*, (2006) *Data Mining* é o processo analítico orientado a explorar dados (em geral grandes quantidades de dados) em busca de padrões consistentes, ou correlações sistemáticas entre variáveis. A descoberta desses padrões ou correlações deve implicar em potencial de utilização ao usuário e devem possuir natureza quantificável. São exemplos de metodologias aplicadas no *Data Mining*: a sumarização, classificação, clusterização, regressão estatística, a serem detalhados no item 2.2.1.

A quinta etapa do KDD corresponde à interpretação dos padrões e correlações obtidas dos dados “minerados” e possivelmente envolve um retorno aos passos da etapa 1 a 4 para a realização de novas iterações. Esta etapa pode envolver a visualização dos padrões extraídos e visualização de dados extraídos dos modelos.

Segundo Freitas *apud* Lemos *et al* (2005), o conhecimento a ser descoberto deve satisfazer a três propriedades: ser correto (tanto quanto possível); ser compreensível pela maioria dos usuários; e ser interessante, útil, novo e surpreendente. Ainda segundo o autor, o método de descoberta do conhecimento deve apresentar as seguintes características: ser eficiente, genérico (ou seja, aplicável a vários tipos de dados) e flexível (facilmente modificável). Uma vez que se tenha obtidos dados com características dessa natureza, só então a aplicação dos resultados a um novo conjunto de dados é passível de ocorrer.

Assim, é notória a influência deste campo de estudo as atividades de crédito no setor financeiro. Uma vez que o acesso a um elevado número de informações sobre mutuários é possível, tais como *bureaus* de informação, comportamento das contas correntes, informações sobre o perfil dos mutuários bons e maus pagadores, etc., com a utilização de técnicas de *Data Mining* as instituições financeiras são

capazes de extrair informações relevantes para a determinação do risco de Crédito decisões de concessão de crédito, com alto grau de agilidade e confiabilidade.

A utilização desse ferramental revolucionou a setor bancário em termos de agilidade e eficácia na concessão de empréstimos. Santos (2003) destaca com a aceleração do crescimento da informática, a partir dos anos 70, a abordagem estatística baseada na pontuação de propostas de crédito surgiu como um dos métodos mais importantes de suporte a tomada de decisão para grandes volumes de propostas de crédito para pessoas físicas e jurídicas.

Portanto, em um contexto onde a disponibilidade de informações sobre empresas e indivíduos é de fácil armazenamento e a busca e consulta desses dados é altamente ágil, é natural que as ferramentas do *Data Mining* sejam inseridas no processo de análise de crédito.

2.2.1 Métodos de Data Mining

Fayyad *et al* (1996) apontam que os dois principais objetivos do Data Mining, em termos práticos, tendem a ser a sua utilização para fins de Previsão e de Descrição. O processo de Previsão envolve o uso de variáveis ou campos existentes em bases de dados em uma tentativa de antecipar resultados futuros acerca das variáveis de interesse. Já o processo de Descrição tem como foco a identificação de padrões que de algum modo estejam “escondidos” nos dados devido as limitações humanas de intepretação. Para tanto, os objetivos de Previsão e Descrição podem ser alcançados utilizando-se de uma série de métodos de *Data Mining*.

Nessa linha, Larose (2005) expõe que os principais métodos de Data Mining consistem na *Descrição, Agrupamento de Dados, Associação. Classificação* e *Regressão Estatística.*

Em algumas situações, os pesquisadores apenas buscam descrever padrões ou tendências que possam existir nos dados. Para tanto, o método da *Descrição* busca apresentar os dados de uma determinada base de modo o mais transparente possível, de forma que a interpretação seja intuitiva e conhecimentos possam ser extraídos.

O *Agrupamento de Dados*, também referenciado como *Clustering,* tem como objetivo identificar e aproximar os registros similares. Um agrupamento (*Cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Este método não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares.

O método da *Associação* consiste em identificar quais atributos estão relacionados. Apresentam a forma, “SE atributo X ENTÃO atributo Y”. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises de afinidade utilizadas na identificação de "cestas de compras”, onde identificamos quais produtos são levados juntos pelos consumidores.

A *Classificação* é o método de mineração de dados mais comumente aplicado, e que emprega um conjunto de exemplos pré-classificados para desenvolver um modelo que pode classificar a população de registros em geral. Por exemplo, um modelo de classificação que prediz o risco de desenvolvimento de determinada doença poderia ser desenvolvido com base nos dados observados para muitos pacientes que desenvolveram essa doença ao longo de um período de tempo. Os dados podem ser constituídos pelo histórico de doenças prévias dos pacientes, idade, seu estilo de vida, pressão arterial, resultados de exames etc. Os métodos de classificação poderiam gerar um modelo capaz de identificar um paciente em um grupo de risco com base em suas características. Os métodos de *Classificação* são de natureza discreta e não implicam ordem. Valores contínuos, de natureza numérica, utilizam o método referenciado como *Regressão Estatística.*

A Regressão Estatística é um método matemático que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explanatórias). As aplicações do método da Regressão Estatística são muitos, tais como: estimar a probabilidade de que um paciente vai sobreviver dado os resultados de um conjunto de testes de diagnóstico, estimar previsão de demanda do consumidor por um produto novo como uma função das despesas de publicidade e dentre eles, estimar a probabilidade de que um determinado cliente seja um bom pagador de seus empréstimos com base na análise de um conjunto de suas características.

No âmbito deste trabalho, o método de *Data Mining* que será utilizado é da Regressão Estatística, com particular ênfase na Regressão Logística. Esse método de regressão é de particular aplicação na área financeira devido a sua capacidade de determinação do nível de probabilidade de eventos binários e não-métricos, baseado nas características de suas variáveis regressoras, conforme será apresentado na sequência.

2.3 ESTATÍSTICA MULTIVARIADA

Os processos de tomada de decisão de Crédito necessariamente englobam a necessidade de levar em conta um grande número de fatores. Como evidenciado, os métodos para a análise desses dados podem variar e, dependendo da abordagem adotada, variáveis relevantes podem não ser levadas em conta afetando assim a qualidade da tomada de decisão, principalmente dentro da abordagem qualitativa. Em contra ponto a essa dificuldade a gestão tradicional passou a utilizar processos de avaliação estatística das informações. Os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: um que trata da estatística que olha as variáveis de maneira isolada, a estatística univariada, e outro que aborda de maneira simultânea múltiplas variáveis em um único relacionamento ou conjunto de relações, a estatística multivariada, e que é de particular interesse a este trabalho, pois e neste conjunto que encontramos os Métodos de Data Mining de Regressão Estatística.

Viali (2005) descreve que o campo da análise multivariada refere-se a todos os métodos estatísticos que analisam simultaneamente múltiplas medidas em cada indivíduo ou objeto sob investigação. Qualquer análise simultânea de mais de duas variáveis pode ser, de certo modo, considerado como análise multivariada. Os diferentes métodos da análise multivariada podem ser divididos em dois grupos: as técnicas do tipo Correlação, que contemplam a Análise de Fatores, Escalonamento Multidimensional, Análise de Correspondência e as técnicas do tipo Regressão que contemplam a Regressão Linear Múltipla, a Análise de Variância Multivariada, a Análise Conjunta, a Correlação Canônica, a Análise Discriminante e a Regressão Logística.

Pode-se dizer que a técnica de Regressão Logística, também referencia como Regressão Logit, de certa maneira corresponde a uma combinação de Regressão Múltipla e da Análise Discriminante. Ela é semelhante à análise de Regressão Múltipla no sentido de que uma ou mais variáveis independentes são utilizadas para prever uma única variável dependente. O que distingue Regressão Logística da Regressão Linear Múltipla é que a variável dependente é não-métrica como na Análise Discriminante. É justamente esta característica que torna o método de Regressão Logística de particular interesse aos processos de Análise de Crédito uma vez que o resultado das concessões é binário e não-métrico (variável dependente), onde busca-se saber as chances de um mutuário ser ou não um bom pagador em caso da realização de um empréstimo com base no conjunto de suas características (variáveis explanatórias).

Muito embora em muitas outras características seja semelhante à Regressão Linear Múltipla, no caso da Regressão Logística a escala não métrica da variável dependente requer uma abordagem diferenciada na estimação e nas hipóteses sobre a distribuição subjacente. (STOCK et al, 2004 e GUJARATI, 2006)

2.3.1 Análise Discriminante

Para a explicação da metodologia da Regressão Logística cabe-se primeiramente explicar o conceito de função discriminante linear que correspondente a principal ferramenta utilizada na abordagem da Análise Discriminante. A função discriminante linear tem por objetivo classificar um indivíduo X a dois ou mais grupos, admitindo-se que X pertença a um deles. Este processo é realizado através de uma "regra de corte" baseada na similaridade entre os indivíduos de determinado grupo. Essa similaridade é medida em função de "m" variáveis (x_1, x_2, x_3, ... , x_m) que dão aos indivíduos suas características passíveis de diferenciação.

SICSU (2010) demonstra que o estabelecimento de uma regra de corte para a classificação dos indivíduos pode ser representada através de uma função linear. A função discriminante linear entre dois grupos (G1 e G2) é uma função linear de p

variáveis discriminadoras X_1, X_2, X_3, ... , X_m, que permite classificar um individuo em um desses dois grupos. Podemos representá-la por:

$$Z = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3 + \cdots + \beta_m . x_m \quad (1)$$

SICSU (2010) aponta que nessa equação β_0 é uma constante, β_1, β_2, β_3, ... β_m são denominados pesos das variáveis, e Z é denominado escore do individuo. Cujas características são x_1, x_2, x_3 .... x_m. Nestes termos, é determinado um valor predefinido a Z_0, sendo que quando $Z \geq Z_0$ o indivíduo portador das características é classificado no grupo G1 e se $Z < Z_0$, ele é classificado no grupo G2.

No caso da Análise de Crédito, pode-se entender que a decisão entre conceder ou não o crédito a determinado indivíduo trata-se de um processo de análise discriminativa na qual o indivíduo é classificado em um determinado grupo (indivíduos com perspectivas aceitáveis de pagamento dos empréstimos ou bons pagadores e indivíduos com perspectivas não aceitáveis de pagamento de pagamento do empréstimo ou maus pagadores) com base nas suas características.

Existem, portanto, inúmeras formas de se determinar uma função discriminante linear, bem como vários critérios para a fixação do Z_0, sendo que, como indicado, este trabalho se utilizará de metodologia da Regressão Logística, apresentada na sequência.

2.3.2 A Regressão Logística

Gujarati (2006) explica que diferentemente dos modelos de regressão clássicos, tais como o Método dos Mínimos Quadrados, em que o resultado Y de uma função é quantitativo, onde o objetivo da análise é estimar o seu valor esperado ou médio, dados os valores dos regressores, os modelos em que o resultado de uma função é qualitativo (neste caso o pertencer ou não a um determinado grupo de indivíduos com características semelhantes) o objetivo da análise de regressão é

encontrar a probabilidade de que algo aconteça, o que muitas vezes o leva a serem também referenciados como modelos estatísticos de probabilidade.

Dentre as metodologias de análise estatística de modelos de probabilidade, o modelo de Regressão Logística, também às vezes referenciado como modelo de Regressão Logit, permite estimar a probabilidade de que um indivíduo X pertença ao grupo G1 (que pode corresponder a um grupo ou a um evento), sendo que a probabilidade de que ele pertença ao grupo G2 é 1 menos esse valor. No contexto da analise de crédito, podemos tomar G1 como o grupo dos bons pagadores e G2 como o grupo dos maus pagadores. SICSU (2010) aponta que a regressão logística fundamenta-se pela validade da seguinte relação:

$$Ln\left[\frac{P(bom)}{(1-P(bom)}\right] = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \cdots + \beta_m.x_m \qquad (2)$$

Denotando-se a função linear por Z teremos que

$$Z = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \cdots + \beta_m.x_m \rightarrow Ln\,\frac{[P(bom)]}{[1-P(bom)]} = Z \qquad (3)$$

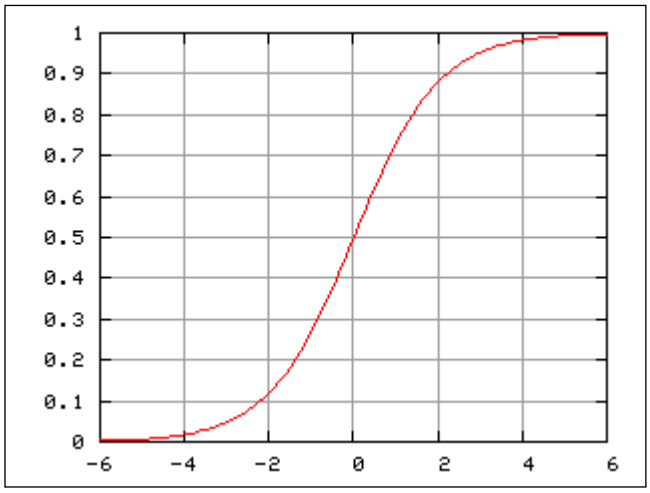
Assim, a probabilidade de um individuo com características x_1, x_2, ... x_m pode ser obtida através da seguinte função[1]:

$$P(bom) = \frac{e^Z}{(1+e^Z)} \qquad (4)$$

A variação entre P(bom) e Z poder ser analisada na Figura 2:

[1] A estimação dos coeficientes da equação ocorre por um processo de máxima verossimilhança. Maiores detalhes sobre o método podem ser consultados em Stock et al (2004), página, 212.

FIGURA 2 - CURVA LOGÍSTICA



Fonte: (SICSU, 2010)

A aplicação junto ao processo de Análise de Crédito ocorre, portanto, pela utilização da metodologia de Regressão Logística dentro da abordagem KDD e *Data Mining*, onde as variáveis x_1, x_2, ... x_m são características dos clientes oriundas de uma determinada base de dados confiável para que seja possível se extrair padrões que permitam estabelecer uma equação preditiva quanto a probabilidade de inadimplência de um indivíduo X. A partir dos valores de Z e de P (bom) obtidos pelas equações (1) e (4), respectivamente, pode-se estabelecer um sistema de *Credit Scoring*[2], no qual, dependendo dos valores obtidos, o cliente será classificado como tendo perspectivas de bom ou mau pagador. No âmbito deste trabalho, a regra de corte para determinar essa classificação será de P (bom) ≥ 0,5 para clientes com perspectivas de serem bons pagadores e de P (bom) < 0,5 para os clientes com perspectivas de serem maus pagadores.

2.3.2.1 Método de Construção do Modelo

[2] SICSU(2010) aponta que os modelos de *Credit Scoring* são uma denominação genérica dada no mercado para as fórmulas de cálculo dos escores de crédito, que possuem a finalidade de quantificar o risco de crédito de maneira objetiva. O autor ressalta que a maneira que essa informação é utilizada no processo de decisão de concessão de crédito varia de acordo com os diferentes gestores de crédito, que atribuem diferentes para o estabelecimento de taxas, garantias, prazos etc. de acordo com os respectivos escores dos clientes.

No presente trabalho, foi utilizado o método de pesquisa empírico-analítico. Esse método está relacionado a uma abordagem prática que envolve a coleta, o tratamento e análise dos dados com o intuito de investigar a existência de relações causais entre as variáveis em estudo. Os dados analisados possuem natureza predominantemente quantitativa e, por isso, técnicas estatísticas foram empregadas nas mensurações. A validação científica de estudos desenvolvidos mediante a aplicação desse método ocorre por meio de testes de significância dos instrumentos utilizados.

Consiste-se, portanto, na construção de modelo através da Regressão Logística através de uma metodologia de identificação de variáveis com poder de explicação. Neste trabalho, inicialmente, todas as variáveis levantadas serão incluídas para construção do modelo, entretanto, no modelo logístico final, apenas algumas variáveis serão selecionadas, sendo que a escolha das variáveis será feita por intermédio do método *Forward*, que é largamente utilizado em modelos de regressão logística. No método *Forward* as variáveis são selecionadas a cada passo, de acordo com critérios que otimizem o modelo. Somente as variáveis realmente importantes para o modelo são selecionadas, e as demais são descartadas. Portanto, para a utilização do método *Forward*, é necessária a definição de critérios para comparação no que diz respeito a qualidade dos modelos. Neste trabalho os critérios selecionados são o Teste Wald para qualidade das variáveis explanatórias, e a Matriz de Confusão como medida de ajuste da Regressão.

2.3.2.2 O Teste WALD

Nomeado em homenagem ao estatístico Abraham Wald o teste Wald é um teste estatístico paramétrico. Semelhante ao teste T aplicado aos modelos lineares, o teste Wald avalia a hipótese nula de que o parâmetro estimado é igual a zero. O parâmetro utilizado no teste Wald é calculado pelo quadrado da razão entre o coeficiente e o seu erro padrão, sendo que a razão resultante, sob as hipóteses H_0 : $\beta_j = 0$ e H_1 : $\beta_j \neq 0$, tem distribuição normal padrão, do onde se extrai seu p-valor.

$$W_j = \frac{\beta_j}{\sigma\beta_j} \quad (5)$$

Define-se então um grau de significância para rejeição da H_0 resultando na escolha de parâmetros independentes estatisticamente diferentes de zero. No caso deste trabalho o grau de significância mínimo para os parâmetros será de p = 0,05, sendo que valores superiores resultam na exclusão das variáveis do modelo.

2.3.2.3 A Matriz de Confusão

Powers (2007) descreve que a Matriz de Confusão (*Confusion Matrix*), é um tipo de tabela que apresenta, de um modo particular, informações sobre as classificações corretas e incorretas realizadas por um sistema de classificação. Essa abordagem é amplamente utilizada para identificar o grau de qualidade dos modelos de Regressão Logística, uma vez que produz uma série de indicadores que expressam diferentes níveis de adequação dos resultados do modelo aos resultados observados, sejam das bases estimadas ou de testes em bases de controle. A tabela na sequência apresenta um exemplo de Matriz de Confusão adequada à realidade deste estudo, onde podem ser sintetizados os resultados de um modelo que estima a probabilidade de dois estados possíveis (neste caso adimplentes e inadimplentes), determinados através da de regra de corte exposta no item 2.3.2:

TABELA 1 - MATRIZ DE CONFUSÃO APLICADA AO CASO DE MODELOS DE ANÁLISE DE CRÉDITO

Resultados	**Classificação do Modelo**	
Classificação Original	Inadimplentes	Adimplentes
Inadimplentes	*a*	*b*
Adimplentes	*c*	*d*

Fonte: Elaboração Própria

- “a” é número de previsões corretas de que um cliente inadimplente é mesmo inadimplente;
- “b” é número de previsões incorretas que clientes inadimplentes são adimplentes;
- “c” é número de previsões incorretas de que bons clientes são inadimplentes;
- “d” é o número de previsões corretas de que um cliente adimplente é de fato adimplente;

Powers (2007) aponta que a partir dos resultados da matriz podem-se obter um conjunto de indicadores de análise do comportamento do modelo, que podem ser de particular interesse ao pesquisador dependendo da finalidade do estudo:

- Acurácia (*AC*) é a proporção de previsões realizadas de modo correto em relação ao total de casos estudados. A Acurácia é determinada pela seguinte equação:

$$AC = \frac{a + d}{a + b + c + d} \quad (6)$$

- Taxa de Positivos Verdadeiros (TP) é a proporção de casos adimplentes que foram corretamente classificados como adimplentes:

$$TP = \frac{d}{c + d} \quad (7)$$

- Taxa de Negativos Verdadeiros (*TN*) é a proporção de casos inadimplentes que foram corretamente classificados como inadimplentes:

$$TN = \frac{a}{a + b} \quad (8)$$

- A Taxa de Falsos Positivos (*FP*) é a proporção de casos inadimplentes que foram incorretamente classificados como adimplentes:

$$FP = \frac{b}{a+b} \quad (9)$$

- A Taxa de Falsos Negativo (*FN*) é a proporção de casos adimplentes que foram incorretamente classificados como inadimplentes:

$$FN = \frac{c}{c+d} \quad (10)$$

3 UM MODELO DE ANÁLISE DE CRÉDITO BASEADO EM DADOS PASSÍVEIS DE SEREM OBTIDOS JUNTO AS BASES DO ESTADO DO PARANÁ

Este capítulo contempla a apresentação do passo a passo adotado até a estimação de um modelo estatístico básico de análise e suporte a decisão de crédito à Fomento Paraná, com foco na utilização de variáveis regressoras cujos valores posteriormente pudessem ser obtidos junto às demais bases de dados administrativas do Governo do Estado do Paraná. Para tanto, foi aplicada a metodologia KDD conforme exposta por Fayyad *et al* (1996), no item 2.3 do Referencial Teórico. A metodologia consiste em um conjunto de procedimentos divididos em 5 etapas, e que correspondem as 5 seções deste capítulo.

Nestes termos, a primeira seção apresenta o processo de seleção de dados, que foi realizado de modo estratégico com o cuidado de que a natureza da informação obtida junto as bases de dados da Fomento Paraná fossem passíveis de serem obtidas nas bases de dados identificadas nas instituições do Governo do Estado do Paraná. A segunda seção corresponde aos principais procedimentos pré-processamento e limpeza dos dados adotados, bem como apresenta o conjunto de estatísticas descritivas referentes aos dados utilizados. A terceira seção relata rapidamente os procedimentos adotados na etapa de transformação dos dados e detalhes sobre as ferramentas computacionais utilizadas. A quarta seção relata a aplicação da metodologia de Data Mining adotada: a Regressão Logística. Finalizando o capítulo, a quinta seção apresenta uma breve interpretação dos resultados obtidos.

3.1 O PROCESSO DE SELEÇÃO DE DADOS

O processo de seleção de dados envolveu duas etapas subsequentes. Primeiramente, investigou-se de modo extensivo quais das instituições que compõe o Governo do Estado do Paraná, pela natureza de sua atuação possuem bases de dados estruturadas e dentre estas, quais delas contém dados que sejam relevantes sob a ótica das teorias de Análise de Crédito. Após a realização desse levantamento

e da estruturação das informações obtidas, iniciou-se o processo de averiguação dos dados existentes nas bases institucionais da Fomento Paraná, de modo e contemplar variáveis que tivessem sido encontradas nas bases de dados do Governo do Estado Paraná. Ambos os processos são detalhados nos subitens a seguir.

3.1.1 As bases de dados do Governo do Estado do Paraná de potencial interesse ao processo de Análise de Crédito

Quando é feita a referência ao Governo do Estado do Paraná de modo geral refere-se ao poder executivo, representado pelo governador, muito embora o Estado seja governado por três poderes que contemplam ainda o poder legislativo, representado pela Assembleia Legislativa do Paraná, e o judiciário, representado pelo Tribunal de Justiça do Estado do Paraná e outros tribunais e juízes. Este trabalho fará referência ao Governo do Estado em sua esfera executiva.

Utilizando-se da metodologia da pesquisa descritiva[3], foi realizado levantamento de dados de modo a elucidar quais as instituições do Governo do Estado do Paraná possuem bases de dados que contém informações de interesse ao processo de análise de crédito. Para tanto, primeiramente foi realizada uma pesquisa prévia junto às instituições que compõe a estrutura do Governo do Estado do Paraná[4].

Através de uma análise da natureza da atividade da instituição e pesquisa junto as mesmas, com base nas premissas dos 5Cs apresentadas no Referencial Teórico, foram elencadas quais delas possuem bases de dados relacionados a pessoas físicas e jurídicas do Paraná devido a execução de suas atividades administrativas. Na sequência, essas informações foram validadas através de entrevistas pessoais e por meio telefônico junto às instituições, que confirmaram a

[3] A literatura existente sobre a Metodologia Científica aponta a pesquisa descritiva como um processo que visa à identificação, registro e análise das características que se relacionam com o fenômeno ou processo. Esse tipo de pesquisa pode ser entendido como um estudo de caso onde, após a coleta de dados, é realizada uma análise das relações entre as variáveis para uma posterior determinação dos efeitos resultantes em uma empresa, sistema de produção ou produto.

[4] O detalhamento da atual estrutura institucional e administrativa do Governo do Estado do Paraná pode ser consultado no ANEXO I deste trabalho.

posse e natureza dos dados que dispõem. Posteriormente, foi elaborado um quadro analítico contendo o conjunto de dados que podem ser de particular interesse ao processo de Análise de Crédito pela Fomento Paraná conforme exposto a seguir:

TABELA 2 - DADOS DAS BASES DO ESTADO DO PARANÁ COM POTENCIAL DE UTILIZAÇÃO NO PROCESSO DE ANÁLISE DE CRÉDITO

Instituição	**Descrição da Informação**	**Dimensão da Análise de Crédito**
Secretaria da Fazenda	Pessoa Jurídica - CNAE - Histórico do Faturamento Líquido - Atrasos nos pagamentos de ICMS - Maior atraso no pagamento de obrigações devidas - Maior saldo em aberto nos últimos 12 meses - Pendências financeiras junto ao Estado do Paraná	Capacidade e Caráter
Secretaria de Segurança Pública e Administração Penitenciária	Pessoa Física (Sócios, Cônjuges e Avalistas) - CPF - RG - Data de Expedição do RG - Orgão Emissor do RG - Local de Nascimento (UF) - Data de Nascimento - Estado Civil - Filiação	Caráter
Junta Comercial do Paraná – JUCEPAR	Pessoa Jurídica - Endereço da Sede - Telefone - Número de sócios - CPF dos Sócios - RG dos Sócios - Setor de Atividade - Tempo de Atividade - Quadro de participação societária - Data de entrada do sócio mais recente - Histórico da composição societária - % de participação de sócios em outras empresas	Capital e Caráter
Companhia de Saneamento do Paraná - SANEPAR	Pessoa Física - Histórico de consumo de água em Reais - Histórico de inadimplência junto à instituição - Comprovação de Residência Pessoa Jurídica - Histórico de consumo de água em Reais - Histórico de inadimplência junto à instituição - Comprovação de Residência	Capacidade e Caráter

Companhia Paranaense de Energia - COPEL	Pessoa Física - Histórico de consumo de energia elétrica em Reais - Histórico de inadimplência junto à instituição - Comprovação de Residência Pessoa Jurídica - Histórico de consumo de energia elétrica em Reais - Histórico de inadimplência junto à instituição - Comprovação de Residência	Capacidade e Caráter

Fonte: Elaboração Própria

3.1.2 Os dados selecionados junto as bases da Fomento Paraná

Regidas pela a Resolução BACEN nº 2828, de 30 de março de 2001 as Agências de Fomento têm como objeto social a concessão de financiamento de capital fixo e de giro associado a projetos na Unidade da Federação onde tenham sede. A mesma resolução aponta que cada Unidade da Federação só pode constituir uma Agência, que no caso paranaense corresponde a Agência de Fomento do Paraná – Fomento Paraná. Com seu funcionamento autorizado pela Lei Estadual nº 11.741, de 1997, a Fomento Paraná é caracteriza-se juridicamente como instituição de economia mista organizada sob a forma de sociedade anônima de capital fechado com capital social majoritariamente pertencente ao Estado do Paraná. Na condição de instituição de natureza pública, sua atuação é naturalmente inclinada ao favorecimento da economia regional, e dentro desse escopo a instituição atua em sintonia com a política estadual de desenvolvimento regional do Paraná, principalmente através de políticas e programas regionais de concessão de crédito.

Dentre as diversas modalidades de crédito operacionalizadas pela instituição uma delas se destaca pela sua escala, vigência e impacto social: o Programa de Microcrédito[5] da Fomento Paraná. Em vigência desde o ano 2000 o Programa de

[5] Existem diversas conceituações de microcrédito. A que Fomento Paraná utiliza é a definida na Lei Federal nº 11.110, de 25 de Abril de 2005 que institui o Programa Nacional de Microcrédito Produtivo Orientado – PNMPO, que o considera-o como o crédito concedido para o atendimento das necessidades financeiras de pessoas físicas e jurídicas empreendedoras de atividades produtivas de pequeno porte, utilizando metodologia baseada no relacionamento direto com os empreendedores no local onde é executada a atividade econômica.

Microcrédito da Fomento Paraná[6] já realizou mais de 50 mil operações, em todas as regiões do Estado do Paraná. Os montantes totais concedidos em crédito pelo programa no final de 2015 somavam mais de R$ 250 milhões.

Como critério de escolha, guiado pelos princípios dos 5Cs, buscou-se utilizar variáveis disponíveis nas bases de dados do programa de Microcrédito da Fomento Paraná que qualificam a situação econômico-financeira dos empreendimentos da amostra, e que ao mesmo tempo fossem variáveis passíveis de serem obtidas junto às bases de dados do Governo do Paraná nos termos da Tabela 2. Ainda, foram utilizadas variáveis que tem como origem o próprio processo de contratação, tais como valor solicitado, prazo para pagamento, renda dos avalistas, adimplência etc. O conjunto de variáveis disponíveis nas bases de dados da instituição selecionadas para análise que atendem a esses critérios foram as seguintes:

- Valor total em Recursos contratado;
- Valor em recursos para Investimento Fixo contratado;
- Valor em recursos para Capital de Giro contratado;
- Prazo Total de Financiamento;
- Faturamento Anual;
- Tempo de Atividade em Meses;
- Número de empregados;
- Soma da Renda dos Avalistas;
- Saldo Vencido até 30 dias;
- Saldo Vencido entre 30 e 60 dias;
- Saldo Vencido entre 60 e 90 dias;
- Saldo Vencido entre 90 e 180 dias;
- Saldo Vencido entre 180 e 360 dias;
- Saldo Vencido superior a 360 dias;

[6] Operacionalizado por uma rede de Agentes de Crédito, o público alvo dos financiamentos do programa de Microcrédito da Fomento Paraná são exclusivamente: pessoas físicas que estão iniciando um empreendimento ou que já exercem uma atividade produtiva, mas ainda não formalizaram o negócio, seja como empreendedor individual ou empresa, que tenham faturamento bruto anual de até R$ 60.000,00 (sessenta mil reais), e que precisem investir para iniciar, melhorar ou ampliar o empreendimento; e pessoas jurídicas que já possuam um negócio formalizado, com faturamento bruto anual de até R$ 360.000,00 (trezentos e sessenta mil reais), e que necessitam de financiamento para melhorar ou ampliar as atividades e a produção. Os valores financiáveis e os prazos de financiamento dependem da finalidade do financiamento e do tempo de atividade do empreendimento, variando entre R$1.000,00 a R$ 15.000,00, e entre 9 meses até 36 meses.

3.2 PROCEDIMENTOS DE PRÉ-PROCESSAMENTO DOS DADOS

A fase inicial do pré-processamento dos dados envolveu a construção de uma base agregando os valores numéricos referentes ao conjunto de variáveis selecionadas para a construção do modelo, de modo a assegurar formatos consistentes de dados e que sejam passíveis de processamento nas próximas etapas. Primeiramente, com o intuito de estruturar uma única variável dependente binária, a ser referenciada como Inadimplência, através de um processo de classificação, foram definidas como pertencentes a classe de variáveis dependentes as seguintes variáveis: Saldo Vencido até 30 dias; Saldo Vencido entre 30 e 60 dias; Saldo Vencido entre 60 e 90 dias; Saldo Vencido entre 90 e 180 dias; Saldo Vencido entre 180 e 360 dias; Saldo Vencido superior a 360 dias.

Com base nesse grupo de variáveis, foi definida a subclasse de indivíduos considerados adimplentes, ou bons pagadores, como os que não apresentaram nenhuma ocorrência de inadimplência durante o curso de seus contratos, ou que apresentaram inadimplência inferior a 60 dias, ou seja, com ocorrências registradas nas variáveis “Saldo Vencido até 30 dias” e “Saldo Vencido entre 30 e 60 dias”. A este conjunto de indivíduos a variável dependente foi valorada como 1.

Para a definição da subclasse de indivíduos considerados como inadimplentes, ou maus pagadores, foram considerados todos aqueles que apresentaram ocorrência de inadimplência superior a 60 dias, ou seja, aqueles indivíduos que apresentaram ocorrências nas variáveis “Saldo Vencido entre 60 e 90 dias”, “Saldo Vencido entre 90 e 180 dias”, “Saldo Vencido entre 180 e 360 dias” e “Saldo Vencido superior a 360 dias”. A este conjunto de indivíduos a variável dependente foi valorada como 0.

Após essa transformação e análise dos dados resultantes, optou-se por escolher um subconjunto do total de operações contratadas junto ao Programa de Microcrédito da Fomento Paraná, que contemplou 2500 operações realizadas durante o período de 01 de Janeiro de 2010 até 31 de dezembro de 2014. Posteriormente esse subconjunto foi dividido em três bases de dados: a base de estimação, a base de controle de adimplentes e a base de controle de inadimplentes.

A base de estimação contemplou 1500 operações realizadas durante o período de 01 de janeiro 2010 a 31 de dezembro de 2013, sendo que 750 foram classificados como bons pagadores e 750 classificados como maus pagadores. As bases de controle de adimplentes e inadimplentes contemplaram operações realizadas durante o período de 01 de janeiro de 2014 e 31 de dezembro de 2014, cada uma contendo 500 adimplentes e 500 inadimplentes respectivamente.

O próximo processo de transformação de dados tratou da exclusão de todos dados referentes os demais aspectos operacionais que são levantados junto aos clientes para a realização da operação de financiamento contidos nos registros, de modo a preservar a identidade dos beneficiários. Optou-se por utilizar uma amostra de empreendimentos excluindo-se os processos que apresentavam dados ausentes para as variáveis explicativas selecionadas. Também se optou pela escolha de operações com o cuidado de se evitar *outliers*, erros de inclusão de dados e inconsistências.

3.3 TRANSFORMAÇÃO DOS DADOS

A ferramenta computacional estatística e econométrica utilizado para a realização da estimação do modelo de análise de crédito através da Regressão Logística foi o Software Gnu Regression, Econometrics and Time-series Library - GRETL. Já o canal de obtenção das informações junto as bases de dados foi o sistema interno de gestão da instituição referenciado como FomentoNet. Nele está contido o conjunto de dados sobre o andamento de todas as operações de crédito realizadas. O formato de alocação é proprietário, porém, é passível de obter o conjunto de dados através de buscas junto ao banco de dados cujos resultados podem ser exportados no formato de arquivo Microsoft EXCEL, que é suportado pelo GRETL, o que facilitou o processo de transformação dos dados.

3.4 *DATA MINING* ATRAVÉS DA APLICAÇÃO DA REGRESSÃO LOGÍSTICA

O primeiro procedimento antes da etapa de estimação do modelo através da aplicação da Regressão Logística é o estudo das estatísticas descritivas acerca das variáveis de estudo selecionadas de onde o comportamento geral das variáveis pode ser observado.

TABELA 3 - ESTATÍSTICAS DESCRITIVAS - PARTE 1

Variáveis	**Média**	**Mediana**	**Mínimo**	**Máximo**
Valor total em Recursos contratado	6.975,50	5.450,00	0,00	15.000,00
Valor em recursos para Investimento Fixo contratado	6.503,10	5.000,00	0,00	15.000,00
Valor em recursos para Capital de Giro contratado	472,38	0,00	0,00	8.346,10
Prazo Total de Financiamento	21,67	24,00	3,00	36,00
Número de empregados	0,89	1,00	0,00	15,00
Faturamento Anual	85.354,00	60.000,00	0,00	360.000,00
Tempo de Atividade em Meses	54,51	48,13	0,00	415,00
Soma da Renda dos Avalistas	4.530,20	3.689,30	724,00	39.584,00
Inadimplência	0,50	1,00	0,00	1,00

FONTE: Elaboração própria

TABELA 4 - ESTATÍSTICAS DESCRITIVAS - PARTE 2

Variáveis	Desvio Padrão	Coeficiente de Variação	Perc. 5%	Perc. 95%
Valor total em Recursos contratado	3.871,70	0,56	2.000,00	15.000,00
Valor em recursos para Investimento Fixo contratado	4.058,60	0,62	0,00	15.000,00
Valor em recursos para Capital de Giro contratado	1.182,20	2,50	0,00	3.000,00
Prazo Total de Financiamento	7,01	0,32	10,00	36,00
Número de empregados	1,37	1,54	0,00	3,00
Faturamento Anual	78.108,00	0,92	850,00	270.000,00
Tempo de Atividade em Meses	41,74	0,77	14,10	137,64
Soma da Renda dos Avalistas	3.098,60	0,68	1.838,60	9.625,70
Inadimplência	0,50	1,00	0,00	1,00

FONTE: Elaboração própria

A partir de então, após a inserção dos dados no software de processamento, iniciou-se o processo de identificação das variáveis com poder explicativo. Como apontado, este trabalho utilizou-se da metodologia *Forward* para a escolha das variáveis que iriam compor o modelo, utilizando como critério de seleção os resultados dos Testes Wald, a 95% de significância estatística, e da Matriz de

Confusão, para análise dos resultados gerais do modelo. A partir deste processo, identificaram-se os resultados apresentados na sequência.

TABELA 5 - RESULTADOS DA REGRESSÃO LOGÍSTICA APLICADA A CARTEIRA DE CRÉDITO DA FOMENTO PARANÁ

Variáveis	**Coeficientes Estimados**	**Erro Padrão**	**Teste Wald**	**P-Valor**
Constante	2,30783E+00	3,56E-01	6,482	9,04E-11
Valor em recursos para Investimento Fixo contratado (VF)	2,51213E-04	3,44E-05	7,295	2,98E-13
Valor em recursos para Capital de Giro contratado (VG)	-6,11073E-04	7,24E-05	-8,441	3,14E-17
Prazo Total de Financiamento (PT)	-2,69049E-01	2,26E-02	-11,91	1,11E-32
Faturamento Anual (FA)	7,17145E-06	1,03E-06	6,963	3,33E-12
Tempo de Atividade em Meses (TA)	9,67025E-03	1,90E-03	5,087	3,65E-07
Soma da Renda dos Avalistas (RA)	2,26744E-04	3,17E-05	7,158	8,18E-13

FONTE: Elaboração própria

A partir dos resultados obtidos podemos escrevê-los em formato de equação. Esta equação ao receber os resultados das variáveis arroladas, produz o resultado Z, que aplicado à equação (4) apresentado no item 2.3.2 fornece a probabilidade de um indivíduo ser adimplente durante o período do contrato de crédito:

$$Z = 2{,}30783 + 0{,}000251213 * VF - 0{,}0006111073 * VG - 0{,}269049 * PT + 0{,}00000717145 * FA + 0{,}00967025 * TA + 0{,}000226744 * RA \quad (11)$$

3.5 INTERPRETAÇÃO DOS RESULTADOS

Esta etapa corresponde ao último passo da metodologia KDD, na qual os resultados são numéricos obtidos são convertidos em conhecimento útil a partir de sua devida interpretação. Com a obtenção da equação (11) resta validar seu potencial explicativo através da análise de seus coeficientes e dos resultados da Matriz de Confusão, processos esses que estão detalhados nos subitens a seguir.

3.5.1 Análise dos Coeficientes

O efeito de cada variável explicativa sobre os escores da operação de crédito pode ser descrito através da análise dos coeficientes. Na sequência segue uma análise individual de cada um deles a partir dos resultados obtidos:

- **Valor em recursos para Investimento Fixo contratado (VF):** O sinal do coeficiente foi positivo, o que revela que indivíduos ou empresas que estejam interessadas em capital para investimentos fixos possuem mais chances de serem adimplentes.
- **Valor em recursos para Capital de Giro contratado (VG):** O sinal do coeficiente é negativo, revelando que, em contraponto aos investimentos fixos, uma maior quantidade de capital de giro no processo de concessão contribui negativamente para as perspectivas de adimplência. Uma hipótese para esse resultado, se analisado em conjunto com os resultados do coeficiente de investimentos fixos pode apontar que empresas que estejam buscando capital de giro apresentem indícios de dificuldades de liquidez e geração de caixa, e que empresas que estejam em busca de recursos para investimentos fixos demonstrem situação diferenciada, almejando a estruturação de seu negócio, e que isso se reflete nas suas perspectivas de pagamento;
- **Prazo Total de Financiamento (PT):** O sinal negativo do coeficiente aponta que empréstimos com maiores períodos de amortização apresentam maiores probabilidade de inadimplência. Este é um resultado que deve ter sua análise aprofundada posteriormente, pois, apesar do fato de que quanto mais tempo

tiverem os contratos, maior será a exposição ao risco, por outro lado, quanto maiores os prazos para pagamentos, as parcelas dos financiamentos tendem ser menores.

- **Faturamento Anual (FA):** O sinal positivo deste coeficiente está em linha com a literatura e aponta que quanto maior a capacidade de geração de vendas da empresa maior será sua capacidade de honrar seus compromissos financeiros.
- **Tempo de Atividade em Meses (TA):** Outro resultado que está em linha com a literatura diz respeito ao tempo de atividade da empresa. O sinal positivo do coeficiente representa que quanto mais tempo de existência a empresa possuir maior será a probabilidade de que ela seja adimplente em um contrato de concessão de crédito. Uma hipótese básica para este comportamento é o fato de que, de modo geral, empresa com um maior tempo de existência apresenta maiores chances de já terem validados seus modelos de negócios, e consequentemente maiores condições de honrar compromissos financeiros assumidos.
- **Soma da Renda dos Avalistas (RA):** O sinal positivo do coeficiente aponta que, de modo geral, as chances que os mutuários sejam adimplentes no curso dos contratos é maior quando a soma da renda dos avalistas da operação é maior. A hipótese básica que explica esse resultado é o fato de que, caso a empresa não venha a cumprir com suas obrigações quanto maior for a renda dos avalistas da operação, maiores serão as chances da empresa continuar adimplente.
- **Número de empregados e Valor total em Recursos contratado:** Ambas as variáveis foram excluídas do modelo final tendo em vista que não passaram no Teste Wald a um nível de significância estatística de 95%.

3.5.2 Resultados da Matriz de Confusão para diferentes bases

O processo de classificação dos clientes está baseado no resultado obtido pela equação (11) que possui como termos dependentes, além do termo constante, um conjunto de variáveis apontados na TABELA 5, a saber: Valor em recursos para Investimento Fixo contratado (VF), Valor em recursos para Capital de Giro contratado (VG), Prazo Total de Financiamento (PT), Faturamento Anual (FA),

Tempo de Atividade em Meses (TA) e Soma da Renda dos Avalistas (RA). Para fins de exemplificação da correta utilização do modelo, tomemos como referência um caso aleatório da base de estimação que figurou dentre o subconjunto dos adimplentes e que, no ato do processo de concessão de crédito, apresentava as seguintes características:

- Investimento Fixo contratado (VF): R$ 5.000,00
- Valor em recursos para Capital de Giro contratado (VG): R$ 0,00
- Prazo Total de Financiamento (PT): 18 meses
- Faturamento Anual (FA): R$ 104.880,00
- Tempo de Atividade em Meses (TA): 53,48 meses
- Soma da Renda dos Avalistas (RA): R$ 4.267,85

Ao classificarmos o cliente a partir da metodologia proposta pelo trabalho, os valores das referidas variáveis devem ser atribuídas em seus campos respectivos na equação (11), que passa a ter a seguinte configuração:

$$\mathrm{Z} = 2{,}30783 + 0{,}000251213 * 5.000 - 0{,}0006111073 * 0 - 0{,}269049 * 18 + 0{,}00000717145 * 104.880 + 0{,}00967025 * 53{,}48 + 0{,}000226744 * 4.267{,}85 \quad (12)$$

Após os cálculos, esse cliente apresentou o seguinte valor Z:

$$\mathrm{Z} = 0{,}95806646 \quad (13)$$

O passo seguinte para finalizar a classificação do cliente é aplicar o resultado do valor Z à equação (4) para obter-se a probabilidade de um indivíduo ser adimplente durante o período do contrato de crédito.

$$P(bom) = \frac{e^{Z}}{(\,1 + e^{Z}\,)} \rightarrow P(bom) = \frac{e^{0{,}95806646}}{(\,1 + e^{0{,}95806646}\,)} \rightarrow P(bom) = 0{,}722 \quad (14)$$

Conforme exposto no trecho final do item 2.3.2, a regra de corte para determinar essa classificação é de P (bom) ≥ 0,5 para clientes com perspectivas de serem bons pagadores e de P (bom) < 0,5 para os clientes com perspectivas de

serem maus pagadores. Como nesse caso o resultado foi de P (bom) = 0,722 o cliente foi classificado pelo modelo como adimplente.

Como se trata de um caso passado no qual sabemos o resultado final da concessão, podemos confrontá-lo com o provável resultado apontado pelo modelo, que nesse caso em específico foi correspondido. Ou seja, se o modelo tivesse sido utilizado como regra de decisão na ocasião, teria apontado uma provável situação de adimplência, o que de fato ocorreu. Ao realizarmos o mesmo processo para todas as situações existentes nas bases de estimação e controle, obteremos a Matriz de Confusão, na qual todos os resultados previstos e realizados são confrontados. A partir desse conjunto de resultados podemos inferir o ajuste do modelo à realidade a partir de um conjunto de índices obtidos a partir da Matriz de Confusão.

Nestes termos, abaixo seguem os resultados da capacidade preditiva do modelo, apresentados através da Matriz de Confusão para as bases de estimação e de controle.

TABELA 6 - ANÁLISE DA CAPACIDADE PREDITIVA DO MODELO - BASE DE ESTIMAÇÃO

	Classificação do Modelo	
Classificação Original	**Inadimplentes**	**Adimplentes**
Inadimplentes	560	189
Adimplentes	167	584

FONTE: Elaboração Própria

- Acurácia (AC) = 0,76
- Taxa de Positivos Verdadeiros (TP) = 0,78
- Taxa de Negativos Verdadeiros (TN) = 0,75
- Taxa de Falsos Positivos (FP) = 0,25
- Taxa de Falsos Negativos (FN) = 0,22

A base de estimação (que, conforme exposto, contemplou 1500 operações do Programa de Microcrédito do Fomento Paraná realizadas durante o período de 01 de janeiro 2010 a 31 de dezembro de 2013, dentre as quais 750 foram

classificadas como adimplentes e 750 classificados como inadimplentes) teve seu comportamento corretamente previsto pelo modelo com um grau de Acurácia na ordem de 76%. Ainda, caso o modelo tivesse sido utilizado na ocasião, teria errado 22% das vezes em prever corretos casos de adimplência e 25% nos casos de corretas inadimplências.

As bases de controle de adimplentes e inadimplentes (que contemplou operações realizadas durante o período de 01 de janeiro de 2014 e 31 de dezembro de 2014, cada uma contendo 500 adimplentes e 500 inadimplentes respectivamente) também tiveram seus resultados confrontados com as previsões realizadas pelo modelo, restando os seguintes resultados:

TABELA 7 - ANÁLISE DA CAPACIDADE PREDITIVA DO MODELO - BASE DE CONTROLE DE ADIMPLENTES

	Classificação do Modelo	
Classificação Original	**Inadimplentes**	**Adimplentes**
Inadimplentes	0	0
Adimplentes	148	352

FONTE: Elaboração Própria

- Acurácia (AC) = 0,70
- Taxa de Positivos Verdadeiros (TP) = 0,70
- Taxa de Negativos Verdadeiros (TN) = N/A
- Taxa de Falsos Positivos (FP) = N/A
- Taxa de Falsos Negativos (FN) = 0,30

TABELA 8 - ANÁLISE DA CAPACIDADE PREDITIVA DO MODELO - BASE DE CONTROLE DE INADIMPLENTES

	Classificação do Modelo	
Classificação Original	**Inadimplentes**	**Adimplentes**
Inadimplentes	410	90
Adimplentes	0	0

FONTE: Elaboração Própria

- Acurácia (AC) = 0,82
- Taxa de Positivos Verdadeiros (TP) = N/A
- Taxa de Negativos Verdadeiros (TN) = 0,82
- Taxa de Falsos Positivos (FP) = 0,18
- Taxa de Falsos Negativos (FN) = N/A

Em linha com os resultados obtidos nas bases de estimação, observa-se que a Acurácia para as bases de controle de adimplentes e inadimplentes apresentou graus de acerto na ordem de 70% e 82%, respectivamente. Esses resultados evidenciam uma característica muito positiva do modelo, revelando que o mesmo, na maior parte das vezes, estará certo ao identificar os maus pagadores.

4 CONCLUSÃO

Sem a intenção de esgotar completamente as possibilidades da abordagem apresentada, este trabalho teve como objetivo estimar um modelo estatístico básico de análise e suporte a decisão de crédito à instituição financeira de desenvolvimento Fomento Paraná, com foco na utilização de variáveis regressoras cujos valores posteriormente pudessem ser obtidos junto às demais bases de dados administrativas do Governo do Estado do Paraná. Através da aplicação de técnicas de *Data Mining* junto às bases de dados histórica de operações de Microcrédito da Fomento Paraná, com particular ênfase na Regressão Logística, extraíram-se padrões de comportamento a respeito de clientes adimplentes e dos que apresentaram inadimplência no curso das operações, chegando-se a um modelo estatístico com Acurácia preditiva de até 76%, apresentada nas bases de estimação e de até 82%, apresentada nas bases de controle, e que cumpriu com o propósito de contemplar variáveis regressoras cujos valores pudessem ser acessados nas bases de dados administrativas do Governo do Estado do Paraná.

Como explicitado, a apresentação de um modelo desta natureza não consiste em um esforço finalizado e sim o primeiro passo de uma iniciativa que visa dar visibilidade as possibilidades que se abrem com a utilização deste tipo de ferramental. Através de sua utilização é possível a busca prévia de mutuários qualificados através do acesso as variáveis das bases de dados do Governo do Estado do Paraná, permitindo que instituição identifique potenciais mutuários ao crédito antes mesmo de ser acionada, fato que pode alterar de modo radical o modelo atual da política pública de financiamento, tornando-a mais eficaz e eficiente, e, em última análise, potencializando as ações do Governo do Estado que poderá, por exemplo, monitorar o comportamento de setores estratégicos e buscar influenciar suas rotas de crescimento das economias regionais de maneira muito mais dinâmica. Unem-se, portanto, através dessa abordagem a tecnologia da informação e as técnicas de Data Mining a serviço do desenvolvimento econômico do Estado.

Ainda, os resultados obtidos apontam para um grande potencial de avanços no processo de Análise de Crédito da Fomento Paraná no que diz respeito a utilização conjunta do ferramental do *Data Mining*, e dos dados de posse do

Governo do Estado do Paraná. A expansão desse tipo de abordagem a demais linhas de ação da instituição, tais como financiamentos com valores superiores aos do microcrédito, por exemplo, são extensões naturais deste trabalho. Também o processo acompanhamento das operações (pós-crédito) pode ser alvo de avanço uma vez que pode ser monitorado o nível de atividade das empresas com contratos vigentes, podendo-se sofisticar muito o processo de previsão de inadimplência com base no nível de faturamento corrente das empresas por exemplo.

Não foi alvo deste trabalho debater a fundo o *compliance* que rege o acesso as informações contidas nas bases de dados administrativas do Governo do Estado do Paraná, mas sem dúvidas, existe todo um conjunto de regras de sigilo que devem ser devidamente atendidas para a utilização dessas informações, mas que não se demonstram como um obstáculo definitivo para aplicação desta abordagem. Ressalta-se que existem exemplos de iniciativas semelhantes em operação em instituições de natureza similar em pelo menos três Estados da Federação, a saber: no Banco de Desenvolvimento de Minas Gerais – BDMG sediado no Estado de Minas Gerais e que foi pioneiro no Brasil neste tipo de abordagem, na Agência de Fomento de São Paulo – DesenvolveSP no Estado de São Paulo e na Agência de Fomento de Góias – Goiás Fomento no Estado de Goiás. Essas iniciativas se encontram em diferentes níveis de desenvolvimento, mas todas apresentam sinais favoráveis após sua implantação, sendo que há relatos de demais instituições seguindo a mesma direção, restando à Fomento Paraná a facilidade de trilhar um caminho previamente validado por demais instituições congêneres.

Dentre os demais pontos positivos oriundos da aplicação desta abordagem podem ser citados: ganhos de agilidade no processo de determinação de crédito; a padronização das decisões uma vez que a utilização de parâmetros matemáticos para a tomada de decisão diminui o espaço das decisões qualitativas e com base critérios mal especificados; avanços na gestão de risco das operações que passam a ser pautadas por parâmetros quantitativos de fácil ajuste de acordo com a evolução dos níveis de inadimplência; otimização dos esforços de venda uma vez que poderão ser determinados previamente os clientes que possuem mérito de crédito nos diversos municípios e regiões do estado; elaboração de ações específicas em setores da economia regional etc.

O modelo apresentado no trabalho corresponde a uma versão inicial, que pode ser mais elaborada com a introdução de dados de demais bases tais como as do Banco Central - SISBACEN, as do SERASA Experian, as do Serviço Federal de Processamento de Dados - SERPRO e demais outras que guardam potencial de melhorar consideravelmente os resultados obtidos. Ainda, o avanço na estruturação de dados oriundos de demais órgãos do Governo do Estado demonstra-se como uma possibilidade, uma vez que a continuidade dos esforços institucionais junto aos órgãos pode revelar a posse de ainda mais dados que podem vir a se demonstrar sensíveis ao processo de analise de crédito. Ainda, o modelo apresentado não tem a pretensão de consistir em uma ferramenta definitiva no processo de concessão de crédito, mas sim de suporte a decisão por parte dos Analistas de Crédito da instituição.

Finalizando, espera-se que os resultados apontados pelo trabalho inspirem ações dessa natureza junto à instituição Fomento Paraná a fim de proporcionar ganhos de eficiência que em última análise são ganhos públicos, sociais, na qual a dinâmica do desenvolvimento econômico sai favorecida. O crédito não é o fator exclusivo determinante do sucesso econômico das regiões, mas uma vez colocado à disposição da sociedade com a motivação correta e guiado pela busca do ganho social, possui indubitavelmente um potencial transformador a favor de todos.

REFERÊNCIAS

BAUER, G. FOLEY, M. LAMONTE, M. **Proposed Bank Rating Methodology** Moody`s Investors Service. Moody`s Rating. Nova Iorque. 2014.

BARONE, F. M. LIMA, P. F, REZNDE, V. **Introdução ao Microcrédito.** Conselho da Comunidade Solidária, Brasília, 2002.

CARVALHO, C. E. TEPASSÊ A. C. **Banco Público como banco comercial e múltiplo: elementos para a análise do caso brasileiro.** Livro Bancos públicos e desenvolvimento. Instituto de Pesquisa Econômica Aplicada – IPEA. Rio de Janeiro, 2010.

CHAKRABARTI, S. **Data Mining Curriculum: A Proposal (Version 1.0)**, Intensive Working Group of ACM SIGKDD Curriculum Committee, 2004

SICSU, A. L. **Credit Scoring: desenvolvimento, implantação, acompanhamento**. Editora Blucher, 2010.

SCHRICKEL, W, K. **Análise de Crédito: concessão e gerência de empréstimos.** 4^0 edição. Editora Atlas. São Paulo. 1998

SANTOS, J. O. **Análise de Crédito: empresas e pessoas físicas.** 2^0 edição. Editora Atlas. São Paulo. 2003

FAYYAD, U. SHAPIRO, G. P. SMYTH, P. **From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.** AI Magazine. 1996

FONSECA, O. L. H. NETO, F. D. M., SOUZA, F. J. **Modelos de análise de crédito utilizando técnicas de aprendizado de máquina.** Universidade do Estado do Rio de Janeiro – UERJ, Rio de Janeiro, 2008.

GUJARATI, D. N. **Econometria Básica.** Rio de Janeiro. Elsevier, 2006

GUIMARÃES, J. B. **Financiamento de Micros e Pequenas Empresas em uma Instituição Pública de Crédito,** Universidade Católica de Minhas Gerais - PUC/MG, Minas Gerais, 2002.

HAND, D. MANILLA, H. SMYTH, P. **Principles of Data Mining.** The MIT Press. Massachusetts. 2001

HOFFMANN, R. **Estatistica Para Economistas 4^0 edição revista e ampliada.** Editora Pioneira. Rio de Janeiro. 2006

KENNEDY, P. **Manual de Econometria. 6° Edição.** Elsevier. Rio de Janeiro. 2009

LAROSE, D. T**. Discovering Knwoledge in Data, An Introduction to Data Mining.** John Wiley & Sons. New Jersey, 2005

LEMOS, E. P. STEINER, M. T. A. NIEVOLA, J. C. **Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining**. Revista Administração, v.40, n.3, p.225-234, São Paulo, 2005

POWERS, D. M. W. **Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation**. School of Informatics and Engineering. Flinders University of South Australia. Adelaide. 2007

STOCK, J. H. WATSON, M. W. **Econometria**. Pearson Education - Addison Wesley. São Paulo. 2004

VIALI, L. **Estatística Multivariada.** Universidade Católica do Rio Grande do Sul. Rio Grande do Sul, 2009.

ANEXO I – ESTRUTURA ADMINISTRATIVA DO GOVERNO DO ESTADO DO PARANÁ

A atual estrutura do Governo do Estado do Paraná contempla as seguintes instituições:

a) Chefia do Poder Executivo:

- Governadoria;
- Vice Governadoria;

b) Secretarias de Estado:

- Secretaria de Administração e Previdência;
- Secretaria de Agricultura e Abastecimento;
- Casa Civil;
- Secretaria de Comunicação Social;
- Secretaria da Cultura;
- Secretaria do Desenvolvimento Urbano;
- Secretaria da Educação;
- Secretaria de Esporte e Turismo;
- Secretaria da Fazenda;
- Secretaria da Infraestrutura e Logística;
- Secretaria de Ciência, Tecnologia e Ensino Superior;
- Secretaria de Justiça;
- Secretaria da Cidadania e Direitos Humanos;
- Secretaria de Meio Ambiente e Recursos Hídricos;
- Secretaria de Planejamento e Coordenação Geral;
- Secretaria de Segurança Pública e Administração Penitenciária;
- Secretaria do Trabalho e Desenvolvimento Social;
- Secretaria da Saúde;

c) Departamentos de Assessoramento:

- Chefia de Gabinete;
- Assessoria de Assuntos Estratégicos;
- Cerimonial;
- Assessoria de Relações Internacionais;

d) Órgãos de apoio técnico:

- Casa Militar;
- Procuradoria Geral do Estado;
- Controladoria Geral do Estado;

e) Órgãos de controle do Estado através da Administração Indireta na qualidade de Autarquias:

- Administração dos Portos d Paranaguá e Antonina – APPA;
- Agência de Defesa Agropecuária do Paraná – ADAPAR;
- Centro Cultural Teatro Guaíra – CCTG;
- Centro Paranaense de Referência em Agroecologia – CPRA;
- Coordenação da Região Metropolitana de Curitiba – COMEC;
- Departamento de Estradas de Rodagem – DER;
- Departamento de Imprensa Oficial do Estado – DIOE;
- Departamento de Trânsito do Paraná – DETRAN;
- Escola de Música e Belas Artes do Paraná – EMBAP;
- Faculdade de Artes do Paraná – FAP;
- Faculdade de Ciências e Letras de Campo Mourão – FECILCAM;
- Faculdade Estadual de Ciências Econômicas de Apucarana – FECEA;
- Faculdade Estadual de Educação, Ciência e Letras de Paranavaí – FAFIPA;
- Faculdade Estadual de Filosofia, Ciências e Letras de União da Vitória – FAFIUV;
- Faculdade Estadual de Filosofia, Ciências e Letras de Paranaguá – FAFIPAR;
- Instituto Agronômico do Paraná – IAPAR;
- Instituto das Águas do Paraná - ÁGUAS PARANÁ;
- Instituto Ambiental do Paraná – IAP;
- Instituto de Pesos e Medidas do Paraná – IPEM;
- Instituto de Terras, Cartografia e Geociência – ITCG;
- Instituto Paranaense de Assistência Técnica e Extensão Rural – EMATER;

- Instituto Paranaense de Desenvolvimento Econômico e Social – IPARDES;
- Junta Comercial do Paraná – JUCEPAR;
- Paraná Esporte;
- Paraná Turismo – PRTUR;
- Rádio e Televisão Educativa do Paraná – RTVE;
- Universidade Estadual de Londrina – UEL;
- Universidade Estadual de Maringá – UEM;
- Universidade Estadual de Ponta Grossa – UEPG;
- Universidade Estadual do Centro-Oeste – UNICENTRO;
- Universidade Estadual do Norte do Paraná – UENP;
- Universidade Estadual do Oeste do Paraná – UNIOESTE;

f) Órgãos de controle do Estado através da Administração Indireta na qualidade de Empresas Públicas:

- Empresa Paranaense de Classificação de Produtos – CLASPAR;
- Instituto de Tecnologia do Paraná – TECPAR;

g) Órgãos de controle do Estado através da Administração Indireta na qualidade de Sociedades de Economia Mista:

- Agência de Fomento do Paraná S/A – FOMENTO PARANÁ;
- Ambiental Paraná Florestas S/A;
- Banco de Desenvolvimento do Paraná S/A - BADEP
- Centrais de Abastecimento do Paraná S/A - CEASA/PR
- Centro de Convenções de Curitiba S/A
- Companhia de Desenvolvimento Agropecuário do Paraná - CODAPAR
- Companhia de Habitação do Paraná - COHAPAR
- Companhia de Informática do Paraná - CELEPAR
- Companhia de Saneamento do Paraná - SANEPAR
- Companhia Paranaense de Energia - COPEL
- Companhia Paranaense de Gás - COMPAGÁS
- Estrada de Ferro Paraná Oeste S/A - FERROESTE
- Minerais do Paraná S/A – MINEROPAR

h) Representações do Estado:

- Banco Regional de Desenvolvimento do Extremo Sul - BRDE
- Instituto de Comércio Exterior do Paraná – CEXPAR

i) Serviço Social Autônomo:

- Ecoparaná
- Paraná Tecnologia
- Paranacidade
- Paranaprevidência
- Paranaeducação